Chemical Science

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2018, **9**, 6922

All publication charges for this article have been paid for by the Royal Society of Chemistry

Combining traditional 2D and modern physical organic-derived descriptors to predict enhanced enantioselectivity for the key aza-Michael conjugate addition in the synthesis of Prevymis™ (letermovir)†

Toni T. Metsänen, ‡[a] Katrina W. Lexa, ‡*[b] Celine B. Santiago,[a] Cheol K. Chung,[c] Yingju Xu,[c] Zhijian Liu,[c] Guy R. Humphrey,[c] Rebecca T. Ruck,[c] Edward C. Sherer[b] and Matthew S. Sigman*[a]

Received 10th May 2018
Accepted 17th July 2018

DOI: 10.1039/c8sc02089b

rsc.li/chemical-science

Quantitative structure–activity relationships have an extensive history for optimizing drug candidates, yet they have only recently been applied in reaction development. In this report, the predictive power of multivariate parameterization has been explored toward the optimization of a catalyst promoting an aza-Michael conjugate addition for the asymmetric synthesis of letermovir. A hybrid approach combining 2D QSAR and modern 3D physical organic parameters performed better than either approach in isolation. Using these predictive models, a series of new catalysts were identified, which catalyzed the reaction to provide the desired product in improved enantioselectivity relative to the parent catalyst.

Open Access Article. Published on 18 July 2018. Downloaded on 8/26/2025 5:25:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Introduction

Catalyst design and optimization for enantioselective syntheses remains largely guided by meticulous and costly empirically-driven experimentation. As such, the development of robust methods for accurately predicting the relationship between catalyst structure and reactivity would have a tremendous impact on reaction development. Since catalyst activity and selectivity are fundamentally tied to structure, one possible approach is to apply quantitative structure–activity relationship (QSAR) modeling, which has traditionally been applied to drug development campaigns.[1] Once identified, QSAR models can rapidly provide activity and selectivity estimates for new structures, thus enabling high-throughput virtual assessment of key structural components for the target of interest. Furthermore, highly weighted descriptors in the QSAR model may indicate structural features that influence the desired activity.[2,3]

QSAR in the context of catalysis remains in its infancy.[4–11] In the reported studies, two distinct approaches have been investigated: (1) the use of traditional QSAR-type descriptors evolved in the medicinal chemistry field, which are often 2D;[4] and (2) the more recent use of parameters derived from quantum mechanics (QM) to accurately model key structural features involved in the hypothesized reaction mechanism.[11] Examples of these descriptor types in the context of catalysis are depicted in Fig. 1. Clearly, these descriptors are innately unique and, if combined, may synergize to provide better statistical fitting of a dataset. Thus, we hypothesized that integrating these

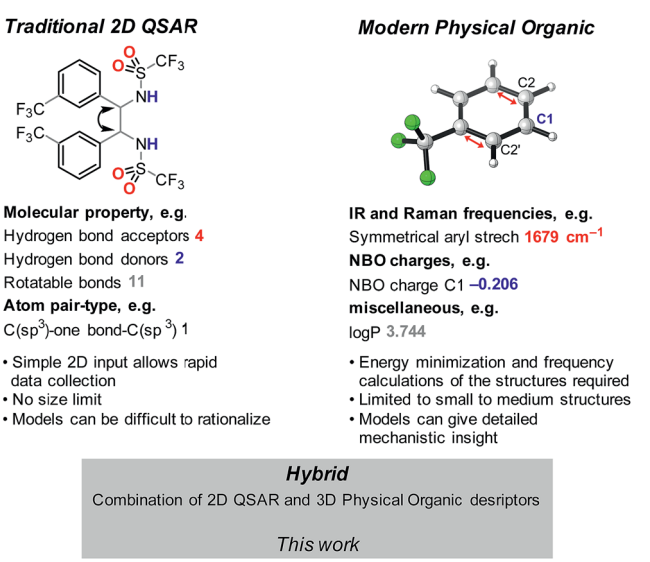


Fig. 1 Descriptors used to establish structure–activity relationships.

[a]Department of Chemistry, University of Utah, 315 South 1400 East, Salt Lake City, Utah 84112, USA. E-mail: sigman@chem.utah.edu

[b]Modeling and Informatics, MRL, Merck Sharp & Dohme, Rahway, New Jersey 07065, USA. E-mail: katrina.lexa@gmail.com

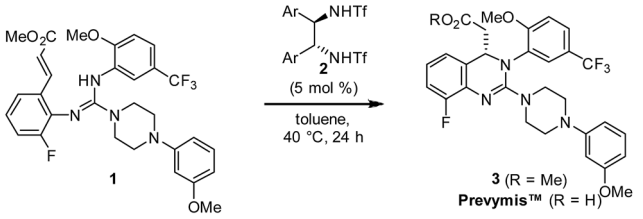
[c]Process Research and Development, MRL, Merck Sharp & Dohme, Rahway, New Jersey 07065, USA

† Electronic supplementary information (ESI) available: Experimental details, computational details and characterization data. See DOI: 10.1039/c8sc02089b

‡ These authors contributed equally.

This journal is © The Royal Society of Chemistry 2018

Open Access Article. Published on 18 July 2018. Downloaded on 8/26/2025 5:25:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.



Training set

2	Ar	%ee(%conv)	2	Ar	%ee(%conv)
a	2,4-F_2-C_6H_3	89.7(100)	o	2-CF_3-C_6H_4	71.4(99)
b	2-F-C_6H_4	89.0(100)	p	2-Me-C_6H_4	69.6(46)
c	2-OTf-C_6H_4	88.3(100)	q	2-pyridyl	64.6(39)
d	4-CN-C_6H_4	87.1(95)	r	2-furyl	58.9(100)
e	Ph	84.7(100)	s	3,5-$(CF_3)_2$-C_6H_3	56.3(99)
f	1-naphthyl	84.5(100)	t	3,5-$(OMe)_2$-C_6H_3	47.9(100)
g	3-CF_3-C_6H_4	83.7(100)	u	2-OAc-C_6H_4	46.3(11)
h	4-Cl-C_6H_4	82.2(100)	v	2-Ph-C_6H_4	39.0(12)
i	2-Cl-C_6H_4	81.9(100)	w	2-OH-C_6H_4	36.2(76)
j	2-Br-C_6H_4	81.3(98)	x	2,4-$(OMe)_2$-C_6H_3	28.3(13)
k	2-I-C_6H_4	80.8(94)	y	2-OMe-C_6H_4	27.5(23)
l	4-OMe-C_6H_4	80.7(99)	z	C_6F_5	23.3(58)
m	2,4-$(CF_3)_2$-C_6H_3	77.7(100)	aa	2,4,6-$(Me)_3$-C_6H_2	15.4(6)
n	3-NO_2-C_6H_4	77.3(100)			

Scheme 1 Asymmetric bistriflamide route for Prevymis™ (letermovir).

two descriptor sets had the potential to improve a recently developed enantioselective aza-Michael conjugate addition, representing a key step in the industrial synthesis of the approved pharmaceutical, Prevymis™ (letermovir; MK-8228).[12,13]

In the initial disclosure, the product enantioselectivity from this key step was 88.3% ee using catalyst **2c** (Scheme 1) for the commercial process. Importantly, as part of the initial optimization effort, a library of 29 bistriflamide catalysts was generated and evaluated under the same reaction conditions with the resultant enantioselectivity (ee) and conversions measured in triplicate. This library provides a wide range of ee's, offering an excellent foundation to examine the potential benefits of QSAR (Scheme 1). Despite the range in ee, this library failed to demonstrate any intuitive SAR trends. For example, during library design, it was hypothesized that amine pK_a would track with ee given the reaction mechanism; however, no relationship was observed. Furthermore, while both electron-withdrawing and -donating substituents (*e.g.*, **2h** and **2l**, respectively) as well as steric bulk (**2f** and **2c**) were tolerated, only four catalysts (**2a–2d**) provided improvement to the parent unsubstituted catalyst **2e**.

On this basis, we set out to identify improved catalysts through the evaluation of QSAR models and hypothesized that the resultant models could provide a more detailed understanding of the salient features required for enantioselective catalysis. Herein, we present a stepwise analysis of both 2D and modern physical organic descriptor-based models as drivers of this optimization campaign. Ultimately, we determined that combining these descriptor sets enabled identification of enhanced catalysts with non-intuitive structural features through a highly predictive virtual screening campaign.

Results and discussion

Training set

In order to accurately model catalyst performance, bistriflamides producing less than 25% conversion were excluded from the training set. Additionally, 2-OH substituted catalyst **2w** was excluded because this catalyst with additional hydrogen bond donating sites may occupy unique transition states. This reduced training set totaled 21 bistriflamides, catalyzing reactions with product ee's ranging from 22.9% to 89.7%. To establish a baseline for the performance of 2D QSAR applied to catalysis, several linear models were built based on method/descriptor combinations with demonstrated successes in the literature and earlier MSD (Merck Sharpe & Dohme) QSAR models related to discovery lead optimization. These include random forest (RF)[14] or support vector machines[15] as the machine learning method, coupled with atom pair substructural descriptors or fingerprint descriptors. In our initial model of the bistriflamide library, we found that RF with Carhart atom pairs (APC)[16] gave the best cross-validated (50/50 split) model of ee, with a cv-R^2 = 0.34 (model **A**, Fig. 2).

As all of the catalysts share a common C_2-symmetric bistriflamide core, we envisaged that the parent arene could be used as an efficient surrogate structure to rapidly explore quantum mechanical-derived physical organic descriptor space. Importantly, a simple arene proxy significantly limited the number of low energy conformations to be considered. Multi-variate linear regression modeling[17,18] produced a relatively simple model using two parameters (NBO_{C2} and P[19]) expressed in three terms (NBO_{C2}, P, and $NBO_{C2} \times P$) to achieve a good statistical fit between the predicted and the measured enantioselectivity for the initial training set of bistriflamides (R^2 = 0.80, LOO-Q^2 = 0.77, model **B**) (Fig. 3). Given the physical organic descriptors identified by our simple arene model, we could interrogate several new structural hypotheses. The unsubstituted C2 NBO charge (NBO_{C2}) represents a general electronic parameter describing the overall electron density of the arene.[20–22] Additionally, the NBO_{C2} term could describe a hydrogen bond interaction between the *ortho* C–H and the substrate. The partition-coefficient (P) describes the hydrophobicity of the arene. This parameter is often used in drug design to predict water solubility of molecules and due to its broad numeric range, is typically reported as log P. We postulated that, in the current case, the partition-coefficient could describe complex attractive interactions between the catalyst and the substrate.

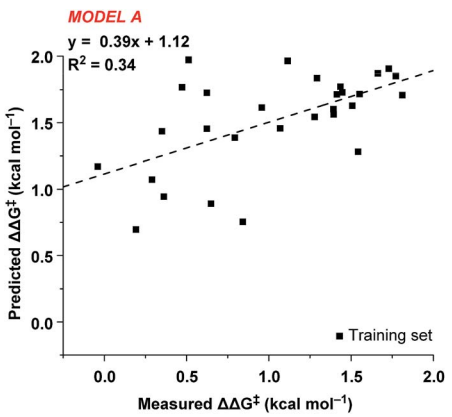


Fig. 2 Cross-validated random forest QSAR model of the training set.

Open Access Article. Published on 18 July 2018. Downloaded on 8/26/2025 5:25:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

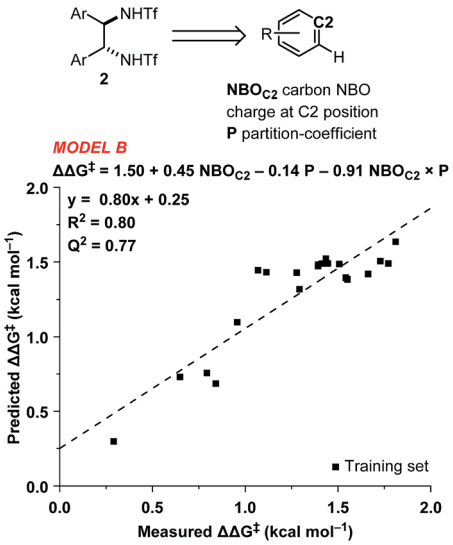


Fig. 3 Modern physical organic descriptors and a predictive model for the initial training set.

Model validation screen: round 1

To validate our initial 2D and 3D models, a set of 30 new catalysts was proposed, 10 of which were synthesized based on predicted product ee (VS1, Table 1, see ESI† for details). Since *ortho*-fluoro substitution seemed a preferred feature, we chose to synthesize several derivatives that incorporated this group, while introducing various *para*-substituents to probe further subtle effects. Additionally, alternative halide- and CF_3-substitution patterns were explored.

Although both models were able to predict enantioselectivity reasonably well, some catalysts were captured better by one of the two QSAR approaches: 2,4- and 3,4-dichlorinated catalysts (**2ae** and **2ai**, respectively) were more accurately predicted by model **B** (3D parameters), while model **A** (2D parameters) gave better predictions for the best three catalysts **2ab–2ad**. It is worth noting that all of the best performing catalysts (**2ab–2ad**) were under-predicted by both models. Overall, the models displayed uneven predictive capability for the poorer performing catalysts (**2ah–2ak**).

Table 1 Virtual screen 1: prospective predictions from the initial models[a]

		Predicted% ee ($\Delta\Delta G^{\ddagger}$)		
2	Ar	Model A	Model B	Measured% ee ($\Delta\Delta G^{\ddagger}$)
ab	2-F-4-Br-C_6H_3	86.7(1.64)	85.9(1.60)	90.2(1.85)
ac	2-F-4-Cl-C_6H_3	87.0(1.66)	85.8(1.60)	89.3(1.79)
ad	2-F-4-CF_3-C_6H_3	87.7(1.70)	83.6(1.50)	88.3(1.73)
ae	2,4-Cl_2-C_6H_3	73.3(1.16)	85.3(1.58)	86.2(1.62)
af	4-CF_3-C_6H_4	84.9(1.56)	81.3(1.43)	85.3(1.58)
ag	4-F-C_6H_4	82.0(1.44)	84.0(1.52)	84.1(1.52)
ah	3-F-C_6H_4	82.8(1.47)	76.6(1.26)	79.9(1.36)
ai	3,4-Cl_2-C_6H_3	69.8(1.07)	81.9(1.44)	78.9(1.33)
aj	2-F-4-CN-C_6H_3	87.1(1.66)	84.6(1.54)	78.5(1.32)
ak	3,4,5-F_3-C_6H_2	62.3(0.91)	66.3(0.99)	75.9(1.24)

[a] $\Delta\Delta G^{\ddagger}$ given in parentheses in kcal mol^{-1}.

Focused virtual screen: round 2

The ten new compounds used as external validations in round 1 were incorporated into our existing dataset to inform and improve our QSAR models. Since the dataset is weighted towards high ee catalysts, we split the data into test and training sets by selecting every other catalyst, starting from the best performing catalyst, until we had a third of the data in a new validation set. The updated 3D QSAR model **B′** was then used to design and predict a new virtual screening set of 35 bistriflamides (VS2, see ESI† for details). The 2D model was updated likewise. Unfortunately, neither the 2D nor the 3D QSAR model predicted any of the new catalysts would give the product in greater than 90% ee. These modest predictions for VS2 could be the result of the inherent difficulty of extrapolating beyond the best identified catalyst (**2ab**, 90.2% ee). In fact, RF is recognized for having a ceiling to predicted activity that is set by the highest activities in the training data.[14] However, given the orthogonal accuracies observed for the 2D and 3D models within VS1, we hypothesized that combining substructural and physical organic descriptors into a hybrid 2D/3D model might be a beneficial strategy. These two model types have been compared, yet very few studies have investigated their potential for synergy,[23–25] and, to the best of our knowledge, none exist in the field of catalyst optimization.

We selected the 20 highest-weighted descriptors from model **A** and added these to the physical organic descriptor set. Both 2D and 3D QSAR parameters were thus available during the process of linear regression modeling.[17] Interestingly, one 2D-QSAR parameter, FX1sp3CX2sp207, stood out as highly synergistic with the 3D physical organic descriptors. The FX1sp3CX2sp207 parameter is a substructural atom pair describing fluorine atoms seven bonds away from an sp^2-hybridized carbon with two non-hydrogen substituents (Fig. 4).

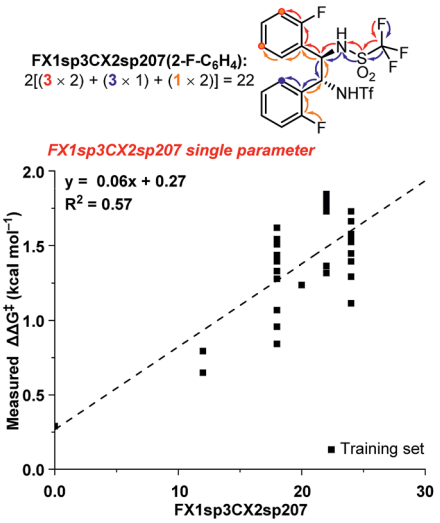


Fig. 4 FX1sp3CX2sp207 single parameter model.

This journal is © The Royal Society of Chemistry 2018

Open Access Article. Published on 18 July 2018. Downloaded on 8/26/2025 5:25:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

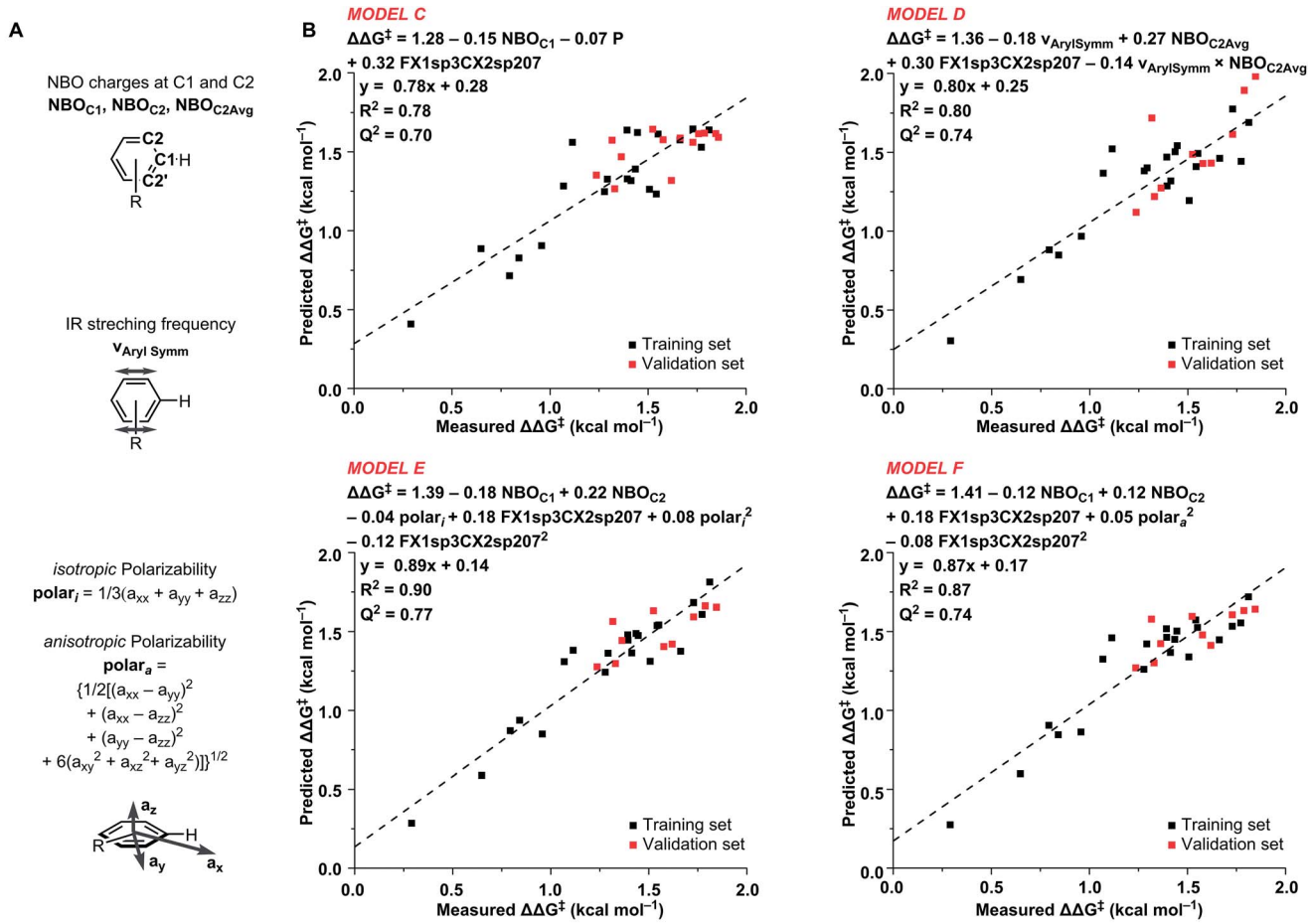
Within the bistriflamide catalyst scaffold, this represents a readout for the unsubstituted *ortho* and *meta* positions and, of particular note, *ortho*- and *meta*-fluorine substitutions contribute as well. This 2D-descriptor was decisively the best single descriptor ($R^2 = 0.62$) and was found to be a significant predictor in nearly all hybrid models. The fundamental drawback of this parameter is the difficulty to extrapolate beyond current structures within the bisaryltriflamide scaffold.

From our model building process, we obtained four new hybrid models (**C**–**F**) in addition to our standalone models **A**′ and **B**′ (Scheme 2C, see ESI† for details). All four new models provided a statistically satisfactory fit to the training set (R^2 = 0.78–0.90 and LOO-Q^2 = 0.70–0.77) and were well-validated against the test set.

Hybrid model **C** incorporates the C1 NBO charge as an electronic parameter and the previously used partition-coefficient together with FX1sp3CX2sp207 parameter. Out of the virtual screening set, model **C** predicted the OCF_3-substituted catalyst **2aq** to be the most selective (94.3% ee).

The second hybrid model **D** combines the symmetrical aryl IR vibrational frequency[26] and the average of C2 NBO charges as physical organic descriptors for the arene ring, and the substructural descriptor FX1sp3CX2sp207. Model **D** accurately predicted the best performing catalyst in the training and validation sets. Prediction of ee in VS2 with model **D** identified four catalysts with potential outputs above 90% ee. In agreement with model **C**, **2aq** was predicted at 95.4% ee. Additionally, catalysts **2al**, **2an**, and **2ao** were predicted to give 90.0%, 92.3%, and 91.2% ee, respectively.

The models **E** and **F** both include C1 and C2 NBO charge, FX1sp3CX2sp207, and a polarizability parameter. Additionally, both models show an excellent fit to the training and validation



C

		predicted %ee(ΔΔG‡)						measured %ee(ΔΔG‡)
		2D QSAR	Phys. Org.	Hybrid				
2	Ar	Model A'	Model B'	Model C	Model D	Model E	Model F	
al	2-F-4-SiMe$_3$-C$_6$H$_3$	82.3(1.45)	83.3(1.49)	85.6(1.59)	90.0(1.83)	95.2(2.30)	85.5(1.86)	90.4(1.86)
am	2-F-4-I-C$_6$H$_3$	86.7(1.64)	85.9(1.60)	86.1(1.61)	92.3(2.00)	88.1(1.72)	88.1(1.76)	88.8(1.76)
an	2-F-4-Bpin-C$_6$H$_3$	81.8(1.43)	82.5(1.46)	85.5(1.59)	86.8(1.65)	99.2(3.41)	86.4(1.66)	87.3(1.68)
ao	2-F-4-SnMe$_3$-C$_6$H$_3$	82.3(1.45)	83.5(1.50)	85.8(1.60)	91.2(1.91)	98.2(2.91)	85.7(1.60)	-
ap	Fc	70.6(1.09)	56.9(0.80)	88.7(1.75)	83.2(1.49)	90.4(1.86)	90.3(1.85)	-
aq	2-OCF_3-C$_6$H$_4$	80.9(1.40)	83.1(1.48)	94.3(2.20)	95.4(2.33)	63.9(0.94)	75.6(1.23)	-

Scheme 2 Physical organic descriptors (A), hybrid models (B), and virtual screen 2 (C); $\Delta\Delta G^{\ddagger}$ given in parentheses in kcal mol^{-1}.

Open Access Article. Published on 18 July 2018. Downloaded on 8/26/2025 5:25:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

sets. The only difference between the two models was the means by which the polarizability was expressed. For model **E**, isotropic polarizability ($polar_i$) was used, which is the calculated average of the *ax*, *ay*, *az* polarizability vectors. In the arene models used in this study, the *az* term was typically small and the isotropic polarizability was effectively the average polarizability along the *xy*-plane. In the model **F**, anisotropic polarizability ($polar_a$) is used instead. Whereas isotropic polarizability can be readily used to compare the polarizability of equally spherical (or planar) molecules, anisotropic polarizability can be used to better describe asymmetrical structures.[27] As expected, where most of the training set consists of flat structures with negligible polarizability along the *z*-axis, the two terms are collinear (Fig. 5). However, the situation changes dramatically when bulkier substituents that increase polarizability along the *z*-axis are incorporated: the 2-OTf substituted catalyst **2c** is the greatest outlier from the training set. Due to this difference in the polarizability term, the predictions between the two models are dramatically different. While isotropic model **E** predicts the silicon (**2al**), tin (**2ao**), and boron (**2an**) substituted catalysts at 95.2–99.2% ee, the anisotropic model **F** predicts them to give only 85.5–86.4% ee. Interestingly, both models **E** and **F** predict the ferrocene catalyst **2ap** at 90% ee.

Based on the predictions from the six models (Scheme 2c), as well as synthetic accessibility, we synthesized three catalysts that were expected to yield high product enantioselectivities: 2-F-4-$SiMe_3$ (**2al**), 2-F-4-I (**2am**), and 2-F-4-Bpin (**2an**). Gratifyingly, the three catalysts all gave full conversion and good ee, with 2-F-4-$SiMe_3$ giving the highest enantioselectivity observed for this reaction to date, 90.4% ee.

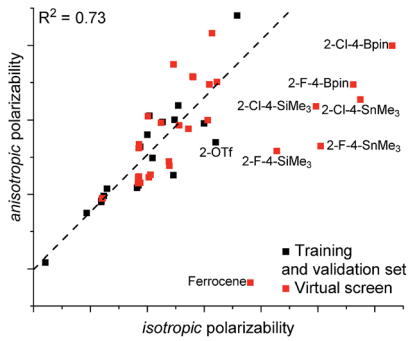


Fig. 5 Correlation between different polarizability terms.

Of the six models used for the virtual screening, both the pure 2D and pure 3D models failed to accurately predict the behavior of the best catalysts. Although models **E** and **F** containing the polarizability terms accurately predicted the iodine-substituted catalyst **2am**, they over- or underestimated the silicon (**2al**) and boron (**2an**) catalysts, respectively. The best effective predictions for VS2 were obtained with the hybrid model **D**.

Focused solvent screening

In our earlier investigation,[13] we had observed that the choice of solvent could have a significant effect on the reaction. Therefore, we performed a focused solvent screen on the best performing catalysts located in this study. While most catalysts performed equally or worse in ethereal solvents compared to the standard toluene conditions (Table 2), the selectivity with best performing catalyst **2al** was further increased giving the highest ee (93.7%) in methyl *tert*-butyl ether (MTBE).

Structural interpretation of the hybrid models

While the combinations of parameters used in the hybrid models do not lend themselves to immediate 3D interpretation, our results made it clear that the strongest-performing models of this reaction require: (1) an electronic parameter, NBO charge or IR frequency; and (2) the substructural parameter FX1sp3CX2sp207. While we anticipated the observed importance of an electronic parameter, its combination with the strong predictive power of the substructural descriptor confirmed our hypothesis that 2D and 3D information could provide orthogonally beneficial information for guiding structure-based design. The atom-pair substructural parameters condense a steric estimate into one numerical value and, in this case, simultaneously amplified the score for fluorine-substituted catalysts. The positive effect of an *ortho*-fluorine substituent is undeniable, but the precise explanation for this effect remains uncertain. We propose that certain *ortho*-substituents, especially fluorine, lock the catalyst in a favorable conformation for the transition state.

Our previous investigation of this reaction mechanism revealed that hydrogen bonding between the catalyst and substrate is a key driver for enantioselectivity.[13] Furthermore, structural modifications of the backbone demonstrated that both sulfonamide N–H's are critical for conversion. DFT

Table 2 Focused solvent screening for selected catalysts[a]

		Measured% ee					
2	Ar	Anhydr. toluene	Wet toluene	MTBE	CPME	EtOAc	2-MeTHF
ab	2-F-4-Br-C_6H_3	90.2	90.2	86.8	88.6	77.8	79.0
al	2-F-4-$SiMe_3$-C_6H_3	90.4	91.8	93.7	92.3	88.9	89.2
am	2-F-4-I-C_6H_3	88.8	88.6	87.1	87.0	78.1	80.5
an	2-F-4-Bpin-C_6H_3	87.3	87.8	89.4	88.7	83.2	84.9

[a] $\Delta\Delta G^{\ddagger}$ given in parentheses in kcal mol^{-1}; MTBE = methyl *tert*-butyl ether; CPME = cyclopentyl methyl ether.

This journal is © The Royal Society of Chemistry 2018

Open Access Article. Published on 18 July 2018. Downloaded on 8/26/2025 5:25:12 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

(density functional theory) calculations with **2c** showed that ring closure most probably occurs *via* aza-Michael conjugate addition; however, we believe that, following substrate tautomerization, a 6-π electrocyclization mechanism can be energetically accessible with some catalysts as well. This structural information from the mechanistic work is consistent with our 2D, 3D, and hybrid models, indicating the utility of performing both machine learning as well as density functional calculations to fully capture the dynamics of a catalyst system.

Conclusions

QSAR offers advantages for catalyst design because well-validated models can facilitate rapid selection and optimization of catalyst properties while lowering labour and resource cost. Here, the best catalyst was found using a hybrid 2D/3D modelling approach. Although only modest gains in enantioselectivity were achieved, it is unlikely that the unusual 2-F-4-$SiMe_3$ catalyst **2al** would have been identified in a traditional intuition-driven screening. This hybrid 2D/3D QSAR modeling strategy should enable guided design and rapid screening of large catalyst libraries, thus facilitating faster route optimization at substantially reduced cost.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors would like to thank Robert Sheridan (MSD) for helpful discussions regarding QSAR strategy and Qunsheng Guo for assistance in catalyst characterization. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. T. T. M. and M. S. S. would like to thank MSD for support. M. S. S. thanks the National Science Foundation (CHE-1361296) for partial support of C. B. S.

References

1 M. L. Drummond and B. G. Sumpter, *Inorg. Chem.*, 2007, **46**, 8613.
2 M. Orlandi, J. A. S. Coelho, M. J. Hilton, F. D. Toste and M. S. Sigman, *J. Am. Chem. Soc.*, 2017, **139**, 6803.
3 M. Orlandi, M. J. Hilton, E. Yamamoto, F. D. Toste and M. S. Sigman, *J. Am. Chem. Soc.*, 2017, **139**, 12688.
4 A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard and A. Tropsha, *J. Med. Chem.*, 2014, **57**, 4977.
5 K. C. Harper and M. S. Sigman, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 2179.
6 S. E. Denmark, N. D. Gould and L. M. Wolf, *J. Org. Chem.*, 2011, **76**, 4337.
7 K. C. Harper and M. S. Sigman, *Science*, 2011, **333**, 1875.
8 K. C. Harper, E. N. Bess and M. S. Sigman, *Nat. Chem.*, 2012, **4**, 366.
9 S. E. Denmark, R. C. Weintraub and N. D. Gould, *J. Am. Chem. Soc.*, 2012, **134**, 13415.
10 C. Yang, E.-G. Zhang, X. Li and J.-P. Cheng, *Angew. Chem., Int. Ed.*, 2016, **55**, 6506.
11 M. S. Sigman, K. C. Harper, E. N. Bess and A. Milo, *Acc. Chem. Res.*, 2016, **49**, 1292.
12 G. R. Humphrey, S. M. Dalby, T. Andreani, B. Xiang, M. R. Luzung, Z. J. Song, M. Shevlin, M. Christensen, K. M. Belyk and D. M. Tschaen, *Org. Process Res. Dev.*, 2016, **20**, 1097.
13 C. K. Chung, Z. Liu, K. W. Lexa, T. Andreani, Y. Xu, Y. Ji, D. A. DiRocco, G. R. Humphrey and R. T. Ruck, *J. Am. Chem. Soc.*, 2017, **139**, 10637.
14 V. Svetnik, A. Liaw, C. Tong, C. J. C. Culberson, R. P. Sheridan and B. P. Feuston, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1947.
15 C. Cortes and V. N. Vapnik, *Mach. Learn.*, 1995, **20**, 273.
16 R. E. Carhart, D. H. Smith and R. Ventkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64.
17 J.-Y. Guo, Y. Minko, C. B. Santiago and M. S. Sigman, *ACS Catal.*, 2017, **7**, 4144.
18 C. B. Santiago, J.-Y. Guo and M. S. Sigman, *Chem. Sci.*, 2018, **9**, 2398.
19 Calculated using Moriguchi's method: I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome and Y. Matsushita, *Chem. Pharm. Bull.*, 1992, **40**, 127, see ESI† for details.
20 F. Weinhold and C. R. Landis, *Discovering Chemistry with Natural Bond Orbitals*, John Wiley & Sons, Hoboken, NJ, 2012.
21 C. A. Hollingsworth, P. G. Seybold and C. M. Hadad, *Int. J. Quantum Chem.*, 2002, **90**, 1396.
22 C. B. Santiago, A. Milo and M. S. Sigman, *J. Am. Chem. Soc.*, 2016, **138**, 13424.
23 H. Matter and T. Pötter, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 1211.
24 E. Estrada, I. Perdomo-López and J. J. Torres-Labandeira, *J. Chem. Inf. Model.*, 2001, **41**, 1561.
25 S. Vilar, E. Estrada, E. Uriarte, L. Santana and Y. Gutierrez, *J. Chem. Inf. Model.*, 2005, **45**, 502.
26 A. Milo, E. N. Bess and M. S. Sigman, *Nature*, 2014, **507**, 210.
27 R. J. W. Le Fèvre, *Adv. Phys. Org. Chem.*, 1965, **3**, 1.